State of Oregon
Department of Environmental Quality Memorandum

Date: November 2, 2005

To: Teresa Michelsen, Avocet Consulting

From: Mike Anderson, Land Quality Division

Cc: Jennifer Peterson, DEQ NWR

Subject: Floating Percentile Calculations

Sorry for the long delay in getting back to you. Other projects and deadlines have intervened. I dug out my old notes and spreadsheets and worked on them again. The comments below may cover some of what is in the memo; however, there are also some new items.

For purposes of this discussion I will refer to how my results compare to the spreadsheet DEQFPPooledCSL.xls that is dated 2/1/2005. I copied all of the columns of data from the ChemData page into my spreadsheet and used that as the starting point for my calculations.

1. As we discussed on the phone, the comment about an agency not being able to select their own %FN came about due to the fact that your instructions say to do the calculations from 5% – 25% and to use the lowest as your starting point for manual calculations. I suspect most people would want to see a variety of solutions in order to compare them but I would also think that when a decision is made it would be to use a specified percentage for the results.

2. I noticed that my spreadsheet counted 254 samples – 129 hits and 125 no hits -- but your spreadsheet indicated there were 258 samples – 129 hits and 129 no-hits. After some searching I found a discrepancy between the BioHits page and the ChemData page on your spreadsheet. The following four fields of data are found on the BioHits page, but <u>not</u> on the ChemData page.

Hit/No-Hit	Survey	Station	Sample
0	MBCREOS4	SED01-12	SED01-12
0	MBCREOS4	SED01-15	SED01-15
0	MBCREOS4	SED01-16	SED01-16
0	MBCREOS4	SED01-30	SED01-30

Since all four of these results are no hit, in your subsequent calculations you also should have only 125 no-hits. So, we apparently used the same number of data points after all.

3. When you create distributions, what is the reason for separating the hit and no-hit data? I have not been able to reproduce your %FP values and, though I can't as yet see why, wonder if my use of a single combined data set may be the reason. I checked my macros but may have

looked at this spreadsheet too often to see errors any more. If you have any suggestions, please let me know.

4. I mentioned on the phone that when I was comparing the results of the two spreadsheets I noticed the following differences. I suspect that these examples are what EPA may be talking about in the bullets on page 3 of the memo. (Note: My three bullets are not meant to correspond to theirs, but I think these points generated their bullets.)

- Some of your final values are lower than mine and have 1 or 2 false positives remaining whereas my results have #FP=0. I assumed you stopped at your value since increasing it would increase the %FN. However, when I increased the result from your value to mine it caused the #FP to go to zero without increasing the %FN. For example, at the 5% level your value for acenaphthene is 1320 and there is 1 FP, whereas my value is 6100 and FP=0. Raising your value to mine reduces FP to 0 without increasing the %FN. On the phone you speculated that it could be because you were not trying to use actual sample data and may have stopped between values due to the tedious nature of the manual calculations. However, when I look at the data distributions I find that both of our values are real data values. You also mentioned that it may be due to using different spreadsheets and having different final numbers, but I get the same results on both spreadsheets.

- Some of your values are higher than mine and both of ours have #FP = 0. Since one of the criteria is to stop when #FP=0 I did not know why you chose to use a higher value and assumed this was because on your spreadsheet this was the point where #FP finally went to zero. However, when I decreased your result to mine, the number of FP remained at zero. For example, at the 5% level your value for acenaphthylene is 640 and mine is 279. Once again these are both actual data points so using interim values is not the reason, and, on both spreadsheets, lowering the value from 640 to 279 maintains #FP=0.

- In some cases where our results differed and either your result or mine had a relatively large number of FPs, your result was often the one that worked better and had lower #FP. These were also the results that were often controlled by %FN.

Out of curiosity I tried to make a better result by combining the two of our results. I used the value that gave the lower #FPs without increasing the %FN. Twelve of your results combined with 25 of mine produced a result with fewer #FPs than either of our original answers. My original set had 181 total FPs, your original set had 179 FPs, and the new set had 148. I am enclosing a spreadsheet that summarizes the results for these three sets. It would be nice to figure out a set process that could give such a result. Although I would prefer to have it all programmed, I agree that this may not be readily achievable.

Even in the combined set of results there are still 6 chemicals that have 10 or more FPs. These are the chemicals for which the results tend to be controlled by the %FN. The values tend to be relatively low compared to the results for many of the other chemicals.

As we have both seen, the results for a large number of the chemicals in the data set end up being controlled by the #FPs. In other words, assuming that I am using the correct terminology, most of the values are simply AETs. As time allows, I would like to run some tests without any of the chemicals that are AETs.

Finally, the problem that still bothers me the most is the one where the result will go up and then drop again as the acceptable %FN increases. This tends to occur when the control of the result switches from being controlled by #FPs at one level and then %FN at the next level. I'm not sure what the best way is to deal with this.

In closing I would say that most of the items on this list do not appear to be deal-breakers, but I would still like to find a reasonably consistent set of instructions for people to use or develop a spreadsheet that we feel can do it all. Also, I think that it would be very helpful to see a comparison of process and results for the Floating Percentile method and the Logistic Regression method.

Enc. Spreadsheet: Comparison of Results.xls